The Oklahoma PetaStore: A Business Model for Big Data on a Small Budget

Patrick Calhoun
University of Oklahoma
350 David L. Boren Blvd.
Norman OK 73019
405-325-4210
phineas@ou.edu

David Akin
University of Oklahoma
350 David L. Boren Blvd.
Norman OK 73019
405-325-5272
david.akin@ou.edu

Joshua Alexander
University of Oklahoma
350 David L. Boren Blvd.
Norman OK 73019
405-325-6417
jalexander@ou.edu

Brett Zimmerman
University of Oklahoma
350 David L. Boren Blvd.
Norman OK 73019
405-325-7176
zim@ou.edu

Fred Keller
University of Oklahoma
2450 John Saxon Blvd.
Norman OK 73071
405-325-6880
fkeller@ou.edu

Brandon George
University of Oklahoma
201 David L. Boren Blvd.
Norman OK 73019
405-325-5113
bcg@ou.edu

Henry Neeman
University of Oklahoma
350 David L. Boren Blvd.
Norman OK 73019
405-325-5386
hneeman@ou.edu

ABSTRACT
In the era of Big Data, research productivity can be highly sensitive to the availability of large scale, long term archival storage. Unfortunately, many mass storage systems are prohibitively expensive at scales appropriate for individual institutions rather than for national centers. Furthermore, a key issue is the set of circumstances under which researchers can, and are willing to, adopt a centralized technology that, in a pure cost recovery model, might be, or might appear to be, more expensive than what the research teams could build on their own. This paper examines a business model that addresses these concerns in a comprehensive manner, distributing the costs among a funding agency, the institution and the research teams, thereby reducing the challenges faced by each.

Categories and Subject Descriptors
B.3.2 [**Design Styles**]: *Mass storage*

General Terms
Design, Economics, Reliability

XSEDE '14, July 13 - 18 2014, Atlanta, GA, USA

ACM 978-1-4503-2893-7/14/07…$15.00.
http://dx.doi.org/10.1145/2616498.2616548

Keywords
Archival storage, mass store, business model

1. INTRODUCTION
In the era of Big Data, and especially as research data management requirements are tightening, research productivity in many disciplines can be highly sensitive to the availability of large scale, long term storage sufficient to contain the many and varied datasets produced and/or consumed by research teams.

At the University of Oklahoma (OU), the OU Supercomputing Center for Education & Research (OSCER), a division of OU Information Technology (IT), has been providing large scale archival storage to a growing population of researchers. This has been accomplished via a resource named the Oklahoma PetaStore, funded by a National Science Foundation (NSF) Major Research Instrumentation (MRI) grant (see Acknowledgements) and consisting of disk and tape hardware, software and media. By adopting an unusual business model, OSCER has made very large scale, long term storage available to researchers, at pricing substantially lower than could be accomplished on their own, and with management provided by IT professionals rather than by research team members (for example, graduate students).

1.1 Strategy
The Oklahoma PetaStore was designed via several key strategies:

- The PetaStore project should distribute the costs of large scale archival storage among multiple entities – grant, institution and research teams – to reduce the cost to each.
- The PetaStore should be an archive (see section 1.3, below), not a backup system and not live storage for research jobs, to reduce the disk I/O transaction load on the PetaStore disk array, and to encourage users to think carefully about what data to archive, and in what manner.

- The PetaStore should be an independent, stand-alone resource, instead of part of a High Performance Computing (HPC) cluster (that is, offline instead of nearline), to allow the PetaStore to be useful not only to HPC cluster users but also to non-traditional users interested in file archiving only, but not in HPC.
- The NSF MRI grant funds should be used to maximize the number of storage media slots, rather than the amount of storage media capacity, by purchasing far more media slots than media, in order to allow the expansion of the resource far beyond the grant budget.
- In order to maximize the value of the PetaStore to its constituency, media slots (vacant bays for disk drives or tape cartridges) should be available first-come, first serve.
- The cost of software should be a modest fraction of total costs, even if this leads to a modest decrease in convenience for the researchers.
- Under the aegis of the OneOklahoma Cyberinfrastructure Initiative (OneOCII) [1], the PetaStore is to be provided to researchers statewide, not just at OU.
- Files stored on the PetaStore should have maximal longevity, for the benefit of research teams, by eliminating ongoing recurring costs and minimizing later incremental costs for retaining datasets over the long term. This requires a clear plan for the follow-on PetaStore II that allows maximal lifespan for tape cartridge media (see section 5, below).

1.2 Motivation

The PetaStore approach has been to distinguish carefully between technology issues and psychosocial issues, providing technology solutions for the technology issues and psychosocial solutions for the psychosocial issues.

For example, the following reasons clarify why some research teams may be hesitant to participate in a large scale, centralized solution:

- Territoriality: There are circumstances under which research teams may be constrained to having their own resources, because of the possibility that a central IT organization may face challenges in providing large scale centralized resources while at the same time serving each user's specific needs well and at high priority.
- Affordability: In certain contexts, storage resources at the research team scale may appear to be less costly than centralized systems, especially in cases where: (a) during the storage design phase, researchers may lack information about the costs associated with space, power and/or cooling, which may not be available at the scale needed, or may only be available at a significant charge (for example, if the scale of the storage resource exceeds the capacity of a wall outlet in an office or produces excessive heat relative to office air conditioning capacity); (b) researcher (especially student) labor can be inexpensive, if research teams expect to have their own team members (especially students) execute that labor, in which case the opportunity cost of lost productivity on domain research objectives may not be clear; (c) the cost of out-year maintenance on, and ultimately replacement of, research group storage resources may not be known or knowable in advance, for example if the scale of the research group's storage resource isn't sufficient to merit a formal invitation to bid/request for proposals process that would allow contractual terms that include a commitment on out-year maintenance pricing as part of each vendor's bid response.
- Data Management: Researchers take a broad variety of approaches to long term management of their data, a full spectrum from long term residence at national repositories to a single disk drive inside the PC of a single graduate student. At the latter extreme, the data in some cases may be rarely or never backed up, and thus vulnerable to a disk drive crash. In addition, when that student graduates, even if the data remains physically present, the rest of the research team may lack knowledge of where the files are, and/or how to read the files (e.g., due to undocumented and/or nonstandard data formats), and/or how to interpret the values in the context in which the datasets were generated and/or analyzed.
- Data Longevity: Institutional ability to make long term commitments to maintain centralized cyberinfrastructure resources such as archival storage can vary broadly, along the full spectrum from permanent to highly temporary. At the latter extreme, research datasets may evaporate, or research teams may suddenly need a substantial amount of capital funding for new archival storage solutions, for example if an institution decides to stop providing the centralized resource. In addition, a centralized resource that is provided long term but on a fee-for-service basis can be problematic. This is especially true for a service with recurring charges (e.g., monthly), in which case research teams may need to continue paying these recurring charges even after the end of the grant that precipitated generation or use of the data, and especially during any funding fallow periods that the research team may experience. Such funding gaps are unusual at some institutions but commonplace at others, and may become an increasing challenge if research budgets continue to tighten, especially at the federal level. Or, a research team may need to fund the recurring charges via a follow-on grant, which may be unrelated to the specific datasets to be funded, in which case expending the later grant's funds on such datasets may be difficult to justify in the later grant's proposal.

These issues arise for the following reasons:

- Institutional Administrations: At some institutions, administrations (including but not limited to central IT organizations) may be perceived by some researchers as barriers to progress rather than partners, perhaps because of some combination of experience and/or anecdotal data.
- Computing Background: Desktop and laptop PCs, and handheld tablets and phones, typically are relatively straightforward to manage, with tiny capital, labor and expertise cost (e.g., handheld storage is typically increased by inserting a MicroSD card, with little or no configuration or deployment labor or expertise required; handheld software is typically installed with a few taps, for a few dollars or free). Such devices, especially handhelds, are increasingly common; for example, 2013 was the first calendar year that over a billion smartphones were sold worldwide, compared to 725.3M in 2012 and 494.4M in 2011 [2]. Thus, a large scale shared resource may be a somewhat alien concept.
- Faculty incentives: At research-intensive institutions where faculty incentives focus primarily on obtaining external funding, publishing, and graduating students, the tradeoffs associated with large scale, shared, centralized resources may appear to favor resources at the research group scale. That is, none of those faculty incentives are directly advanced by expending scarce research funds on computing and/or storage, let alone by having those resources well configured, well managed, secure and reliable (though such incentives can be indirectly advanced by such means).

Ultimately, a key question is, why can't researchers resolve their research data archiving needs by simply purchasing USB disk drives at discount retailers?

For research teams with very small data collections, this may be a perfectly reasonable choice, especially if they're willing to keep multiple copies of mission-critical files on multiple disk drives.

However, beyond a handful of USB disk drives – at the present, perhaps on the order of 10 TB of total data footprint per copy – this approach can become unwieldy to manage, especially for research teams with many different datasets from diverse experiment types. At such a scale, the most likely team-internal solution would be a RAID (or, even better, a pair of RAIDs, for primary and secondary copies), which would require substantially more expertise and labor to deploy and manage, and which would incur significant cost increases, for the RAID enclosure(s), or to purchase server(s)/workstation(s) with many disk drive bays. In addition, team members would have to comply with creation of secondary copies – or the process would have to be automated, requiring even more labor and expertise (though not necessarily capital expense, given the availability of free, open source software products that address such needs).

In addition, USB disk drives (which typically are SATA) are subject to unrecoverable read errors, known informally as "bit rot." In particular, the typical consumer-class SATA bit rot rate is one unrecoverable read error per 10^{14} bits [3], and a 4 TB drive has a total of approximately 32 trillion (3.2×10^{13}) bits, so assuming that individual unrecoverable read errors are independent, then the probability of bit rot is approximately 27% at one full traversal of a 4 TB SATA disk drive, 47% at 2 full traversals, 62% at 3, 72% at 4, 80% at 5, 85% at 6, 89% at 7, 92% at 8, 94% at 9, 96% at 10, and so on. Thus, over the lifetime of a USB disk drive, loss of data has considerable probability. By contrast, LTO tape cartridges have a bit rot rate of one per 10^{17} bits [4] (i.e., 3 orders of magnitude lower probability than consumer-class SATA disk drives), so the probability of data loss during even a significantly extended lifetime is substantially lower.

Thus, a large scale centralized resource such as the PetaStore is best positioned to attract research teams to use it if at least the tape option is cheaper, more reliable, less labor-intensive, reasonably intuitive, and ideally faster (in at least some senses) than a research team's homegrown options (the disk option would be likewise, except for cost).

1.3 Terminology

In this paper, the term *backup* refers to the practice of making automatic daily (or more typically nightly) *incremental* copies of all files that are either new or modified over the past 24 hours, and less frequent automatic *full dumps* of all files regardless of age (for example, every week or every month). Backups are typically characterized by retaining multiple versions of files that have been changed over some period of time, and typically are only accessed in one of the following cases: (a) loss or corruption of one or more files on the storage resource being backed up (for example, if the relevant disk filesystem crashes), (b) accidental deletion of such file(s) by the owner, or (c) dissatisfaction by the owner with the most recent version of a file (a version control system such as git [5] is better optimized for this last case, but not all researcher data owners use version control systems).

For example, in a backup system that performs nightly incrementals and monthly full dumps, a file that was generated 5 years ago and never modified since then will have been written 61 times (the initial write and then 12 full dumps per year for 5 years), and most likely will never have been retrieved, or perhaps will have been retrieved once or a few times.

By contrast, in this paper the term *archive* is used to refer to "Write Once, Read Seldom if Ever." In many cases (including the PetaStore), archiving isn't automatic, but rather is an explicit choice by the user. Deletions likewise are decided by the user (though some archival systems, typically in enterprise rather than research contexts, impose a fixed number of years, for example based on legal mandates with respect to business data retention). Thus an archived file that is 5 years old may have been copied from the original on disk literally just once (though since then it may have been automatically copied from one tape cartridge to another at the archiving software's discretion), and may have been retrieved many times, or a few, or never.

2. BUSINESS MODEL

The PetaStore's business model represents a marriage of external funding and internal funding, from multiple sources. Specifically:

- Grant: Most of the hardware, software and the initial period of maintenance, and a modest amount of media, were funded by a National Science Foundation (NSF) Major Research Instrumentation (MRI) grant (see Acknowledgements).
- Institution: Space, power, cooling, labor and maintenance after the initial maintenance period are being funded by OU, specifically by OU's Chief Information Officer (CIO) and Vice President for Research (VPR).
- Researchers: Storage media (tape cartridges and disk drives) are purchased by research teams, using their own funds, primarily though not exclusively via external funding.

This approach allows the PetaStore to be far more extensible than would be possible if a significant fraction of the grant funds were spent on storage media. In particular, the grant funds have been spent primarily on media slots rather than on media, thereby providing a mechanism by which the storage footprint of the PetaStore can become substantially larger than would otherwise be possible.

Example alternative approaches include: (a) one-time full or partial cost recovery of institutional investments (space, power, cooling, labor, out-year maintenance), for example by applying an upcharge to the cost of each tape cartridge; (b) recurring (e.g., monthly) capacity and/or usage charges, either in addition to or instead of media purchase costs; (c) external cloud storage (e.g., Amazon Glacier [6]); (d) no centralized resource, only research group resources, perhaps with recurring charges for space, power, cooling etc. The disadvantages of these models with respect to the stated context are (i) a potentially higher cost per TB per copy than research group solutions and/or (ii) the need for ongoing funding for recurring costs.

3. IMPLEMENTATION

Because the Oklahoma PetaStore is required to be a robust production resource, the project was never intended to advance the technological state of the art, but rather to leverage existing products and expertise. The hardware choices for the PetaStore are not at all unusual: an IBM DCS9900 [7] disk system (rebranded DataDirect Networks S2A9900) of 1200 disk drive slots, and an IBM TS3500 tape library [8] with 4 LTO-5 tape drives and initially 2859 tape cartridge slots, along with 6 IBM x3650M3 servers [9]. The disk software choice is also commonly used in research computing: IBM's General Parallel File System (GPFS) [10].

However, the tape software choice at the time of purchase was, and remains to date, unusual for this kind of usage: IBM's Tivoli Storage Manager (TSM) [11], which was chosen simply because its price at the time of purchase was quite low as a fraction of total project cost, in comparison to all other tape software (or hardware software combinations) presented to OU in response to OU's PetaStore Request for Proposals (RFP). Thus, the choice of TSM, made for financial reasons, in a sense forced the choice of GPFS and the IBM hardware offerings, because of (a) the high degree of compatibility between these two software products, (b) the fact that IBM provided the only RFP bid response that included TSM, and (c) the fact that IBM's bid response included the TS3500 as the only tape library option and rebranded DDN products as the only disk array options, of which the DCS9900 had the best ratio of cost to number of disk drive slots.

3.1 Hardware

3.1.1 Disk Hardware

The IBM DCS9900 is a large scale disk system consisting of 1200 disk drive slots (20 enclosures of 60 disk drive slots per enclosure), of which 300 slots were initially populated (with 2 TB SATA drives) at time of purchase, with an additional 230 populated since then (also with 2 TB SATA drives), for a total of 530 disk drives deployed to date (44% of slot capacity, ~840 TB useable). The DCS9900 has dual controllers, allowing for failover in the event of a controller failure. Its slot capacity cannot be expanded.

This product was chosen in large part because the cost of a disk system (excluding disk drives) is significantly dependent on the ratio of controllers to disk drive slots; the DCS9900, having only two controllers but 1200 disk drive slots (600 to 1), had a more favorable ratio than the other options evaluated.

The DCS9900 has a nominal peak speed of ~5.4 GB/sec. Idealized benchmarks of the PetaStore configuration have shown ~4 GB/sec.

3.1.2 Tape Hardware

The IBM TS3500 is a large scale tape library, expandable to a total of 18 cabinets, potentially exceeding 22,600 tape cartridge slots. The current system consists of a controller frame (model L53) with 4 TS2350 LTO-5 tape drives (1.5 TB raw per cartridge, 140 MB/sec raw peak bandwidth per tape drive) and 2 tape cartridge expansion cabinets (model S54). The initial system has 2859 tape cartridge slots, with 100 tape cartridges in the original purchase and a total of 920 in place as of this writing (~1.38 PB raw).

The tape library's expandability makes tape an attractive option, not only because of the lower cost per TB for research teams (compared to the cost of disk) but also because it can be expanded far beyond the expected demand.

The project team has targeted sufficient funds to purchase a pair of LTO-6 tape drives (2.5 TB raw per cartridge, 160 MB/sec raw peak bandwidth per drive) in mid-2014 (LTO-6 became available in late 2012, but the team decided to allow shakeout of, for example, firmware issues, and for price per TB to match or improve on LTO-5). This fact is relevant not only to the current PetaStore but also to future plans (see section 5, below).

3.1.3 Servers

The PetaStore is driven by six IBM x3650M3 servers, of which four control the disk system and two control the tape library (in an active/passive configuration). Each server has dual Intel Xeon "Westmere" E5620 CPUs (quad core, 2.4 GHz, 2 x QPI 5.86 Gigatransfers/sec), 24 GB RAM, dual 300 GB SAS 10K RPM disk drives in RAID1, dual QLogic QLE2562 8 Gbps Fibre Channel dual port cards and single Chelsio 10 Gbps Ethernet (10G) dual port card.

3.1.4 Network

The PetaStore hardware components (disk system, tape library and servers) are all connected to OU's Fibre Channel backbone at 4 or 8 Gbps: eight connections at 8 Gbps each for the disk system, eight connections at 4 Gbps each for the tape library, and two to four connections at 8 Gbps each per server. In addition, the servers are connected to OU's 10G Ethernet backbone at one connection per server. The PetaStore also has a 1 Gbps management network.

The Fibre Channel backbone consists of two redundant physical fabrics, resulting in multiple signal paths between all fiber-connected components of the PetaStore. On the servers, disk Logical Unit Number (LUN) paths are aggregated by Linux DM Multipath, and tape device paths are aggregated within the tape drivers themselves.

The connectivity among the relevant components can be seen in Figure 1.

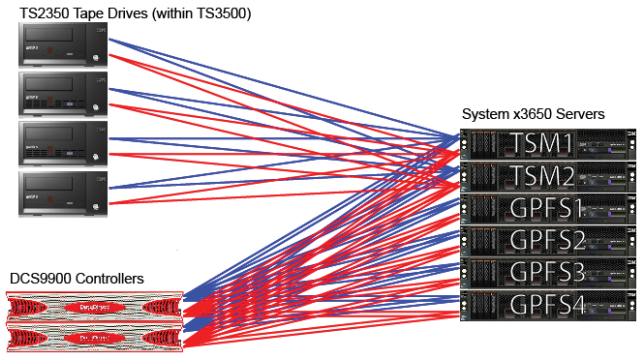

Figure 1: Oklahoma PetaStore SAN configuration (switches not shown).

3.2 Software

3.2.1 Disk Software

The disk system uses GPFS, specifically GPFS Server 3.4 on the four disk system servers and GPFS Client 3.4 on the two tape library servers.

The GPFS Servers are configured such that each disk LUN is owned by exactly one primary and one failover server, manually load-balanced at drive installation time.

Each duplication option (see section 4.3, below) is allocated a GPFS fileset, allowing for efficient and fine-grained replication and migration policies. The default GPFS file placement policy requires redundant replication of both data and metadata blocks for any file staged to the tape system, and no redundant replication of user data targeted for archive on the disk system. GPFS Information Lifecycle Management (ILM) tools link the filesystem to an external storage pool definition for Hierarchical Storage Management (HSM).

GPFS quotas prevent major quota infractions, and also help to enforce the system's minimum file size policy, by setting a maximum number of inodes available to a user, based on their total available media.

3.2.2 Tape Software

The tape software is the aspect of the PetaStore's technology that is the most unusual: Tivoli Storage Manager (TSM).

Specifically, the PetaStore uses TSM Server 6.2 and TSM Space Management client 6.2 on the two tape library servers, and TSM Client 6.2 and TSM Space Management client 6.2 on the four disk system servers. (TSM Space Management is TSM's Hierarchical Storage Management component.)

TSM wasn't specifically designed with research data archiving practices in mind, but rather for backups; in fact, IBM has a product specifically designed for this kind of archiving: High Performance Storage System (HPSS) [12].

This raises a key question: If TSM wasn't particularly designed for this purpose, and HPSS was, why not choose HPSS? The answer is price: HPSS costs several times as much as TSM, and would have consumed the bulk of the project budget, leaving very modest funds for hardware.

In fact, this highlights a key benefit of TSM: The software was purchased with pricing set on a per-host basis, which can be remunerative in backup contexts, where the number of hosts to be backed up can be quite large, but in the PetaStore's configuration, with a total of six hosts, the aggregate price of TSM (and likewise of GPFS) was modest as a fraction of total project budget.

By contrast, common among tape archiving solutions are one or both of the following: (a) an activation charge per tape cartridge slot, and/or (b) a capacity (per-TB) charge attached to the software or hardware. In some cases, these charges can substantially exceed the cost of the tape cartridge that they address, driving up the cost of deploying each tape cartridge by a considerable amount and thereby rendering use of the PetaStore, under the above business model, unaffordable in practice.

In a sense, the technical side of the PetaStore project was fraught with risk: When this project began, the combination of GPFS and TSM was very unusual in large scale archiving, with very few institutions worldwide having adopted this approach. However, this approach is now becoming more popular [13,14,15].

3.3 Configuration Issues

From the user's perspective, file recall from tape is of higher priority than file write to tape, because file writing to tape occurs in the background after the file has been staged to disk. Consequently, the TSM server policies reserve three tape drives for recall, and one tape drive for migrating to tape.

Physical tapes in the PetaStore, though owned by a single entity, can contain data from multiple workgroups. Users are guaranteed their appropriate amount of storage space, but not a specific piece of storage media (i.e., tape cartridge or disk drive). The tape space is maintained through descriptive quotas: maintenance scripts on the TSM server detect minor over-quota infractions and report them to system administrators for resolution.

4. POLICIES AND PROCEDURES

4.1 Use Agreement

PetaStore users must sign a use agreement: no files subject to the Health Insurance Portability and Accountability Act (HIPAA) or the Family Education Rights and Privacy Act (FERPA), nor classified files; for any files subject to an agreement with any Institutional Review Board (at any institution), the user takes full responsibility for compliance; no use by persons not affiliated with a US institution; compliance with OU IT's acceptable use policy; faculty take responsibility for their students' use; they understand the limits of OU's responsibilities.

4.2 Access Mechanisms

The Oklahoma PetaStore currently employs four mechanisms for access:

4.2.1 Cluster Mount

On OSCER's HPC cluster, Boomer, there are two support nodes whose purpose is to allow users to copy data between the PetaStore and Boomer's globally accessible user filesystems (for example, via the cp command). These archive nodes are the only nodes on Boomer that mount the PetaStore filesystem, for multiple reasons:

(a) Cost: Each node that mounts the PetaStore filesystem must have a GPFS client license, which individually is quite inexpensive but which collectively over hundreds of compute nodes can aggregate to a substantial sum.
(b) User behavior: If the PetaStore filesystem were accessible on Boomer on par with other globally accessible user filesystems (for example, /home, /scratch), then the risk would be that some researchers would use the PetaStore filesystem in the same manner as those other filesystems, as a live filesystem for user jobs to output to, which would bog down the PetaStore disk, rendering it less valuable for archiving.
(c) File sizes: If user applications were writing directly to the PetaStore disk, then in practice, some of those applications would write many small files, which is forbidden (see 4.5, below).

4.2.2 Remote Secure FTP/Secure Copy/GridFTP

Except as described below (see 4.2.4), servers outside of OSCER aren't permitted to mount the PetaStore filesystem, for security reasons. Therefore, to store files from and/or to retrieve files to remote systems outside of OSCER, a pair of front end servers, independent of Boomer and of the PetaStore's six servers, are provided. These servers allow Secure FTP (sftp), Secure Copy (scp) and GridFTP access, and mount the PetaStore filesystem, by which mechanism users can access the PetaStore disk space remotely.

4.2.3 Globus Online

OSCER has deployed Globus Online [16] as an additional mechanism for storing and retrieving files from the PetaStore. The primary value of Globus Online is that users aren't required to learn the syntax of GridFTP, which can be challenging, especially for users whose primary computing background is at the level of a desktop/laptop PC or a handheld tablet or phone device (see 1.2, above).

4.2.4 Collocated Servers

OSCER offers an option for users to collocate one or more servers in one of the limited subset of OU's data centers served by OU IT's 8 Gbps Fibre Channel and 10 Gbps Ethernet backbones, so that such a server can directly connect to the PetaStore disk. Such a collocated server needs only a low cost Fibre Channel adapter and a low cost 10 Gbps Ethernet adapter (three to four figures per adapter) to be able to connect to the PetaStore directly. For security reasons, the server would be limited to mounting only the subpartition of the PetaStore disk that is directly relevant to the server's owner(s), on the assumption that there would be a higher risk that a third party server, which might be managed by researchers such as students, could prove to be substantially less secure than a server managed by IT professionals.

4.3 Duplication Options

The duplication option for each file or subdirectory is explicitly chosen by the user, by selecting an appropriate subdirectory within their (or their group's) top-level directory. The options are:

- `/archive/username/disk_1copy_unsafe`

This option is most useful when file retrievals have to be done as quickly as possible and when another copy of each file is stored elsewhere than the PetaStore (for example, at a national center).

- `/archive/username/disk_1copy_tape_1copy`

This option is most useful when file retrievals have to be done as quickly as possible and when the PetaStore is the only place where the files are stored.

- `/archive/username/tape_1copy_unsafe`

This option is most useful when file retrievals don't have to be done quickly and when another copy of each file is stored elsewhere than the PetaStore (for example, at a national center).

- `/archive/username/tape_2copies`

This option is most useful when file retrievals don't have to be done quickly and when the PetaStore is the only place where the files are stored.

Here, `username` is replaced with the specific user's login ID. (In some cases, instead of a username, a project or group name is used.) Note that, for single copy options, the user must explicitly acknowledge the implication of single copy with each file access.

Originally, `disk_2copies` was also considered, but was abandoned based on the relatively modest number of disk drive slots available, coupled with the fact that the disk system cannot be expanded except by purchasing an entire new disk system (at very high initial capital outlay).

The three duplication options that use tape are implemented in different ways from one another: `disk_1copy_tape_1copy` is implemented as a classical backup operation, in that files are copied from disk to tape using the regular TSM backup client, but with the exceptions that previous versions are not maintained, the backup will not expire, and all such copies are "full dumps," that is, not incremental; `tape_1copy_unsafe` is implemented as a periodic HSM migration, leaving a stub file in GPFS for subsequent user retrieval; `tape_2copies` is an identical migration, followed by a periodic tape pool clone operation. All but the most inquisitive users do not distinguish between the stub file from the tape system and a file actually stored on disk, thus helping to simplify user training and PetaStore use.

4.4 Offsite Copies

For files stored under the `disk_1copy_tape_1copy` and `tape_2copies` duplication options, the second copy can be ejected from the tape library and transported to another location. Currently, that location is approximately six miles away, but another site is available approximately 30 miles away. Tape cartridges are migrated every seven days.

4.5 User File Constraints

Tape libraries can have difficulty with large numbers of small files, which aren't desirable, because of strain on the file catalog and the risk of "shoeshining" tape cartridges (excessive small I/O transactions), which can potentially damage the tape media.

One strategy for small files is to confine files below some threshold size to disk only, not to tape. In the case of the PetaStore, however, this approach wouldn't be effective, because users purchase their own media, and if they purchase tape only, then there isn't a pool of unused disk space that can be devoted to their collections of small files.

Instead, on the PetaStore, individual files must be at least on the order of 1 GB per file, with a preference for 10 to 100 GB, but not substantially above 100 GB. Some files are already this size, but for collections of many small files, this requirement can be achieved by creating zip or compressed tar files of the many small files, substantially reducing file count, in some cases by multiple orders of magnitude. Because of the compression associated with zip and tar files, this approach can have the salutary side effect of substantially reducing the footprint of the data being archived.

(Note that the PetaStore doesn't use automatic compression when transferring files from disk to tape, because of a strong preference for compression of files not only on tape but also on disk, combined with a suspicion that a second compression would have little or no positive effect but might increase file transfer times.)

The preferred file size of 10 to 100 GB was chosen as a means of balancing user convenience with performance. In particular, to retrieve a specific file from tape, the aggregate fixed time cost of preparing to retrieve that file from tape includes: (a) if necessary, a tape drive fully rewinds and ejects a previously used tape cartridge that is no longer needed; (b) if necessary, the tape picker robot transports the ejected tape cartridge to an unoccupied tape cartridge slot; (c) the tape picker robot selects the correct tape cartridge and transports it to, and inserts it into, an unoccupied tape drive; (d) the tape drive reads tape identification metadata from the beginning of the tape; (e) the tape drive seeks the file being requested by fast-forwarding that tape cartridge to the location of that file. These aggregate components can vary from tens of seconds to a few minutes, and don't include the time to draw the file down from the tape cartridge after preparation.

Given an LTO-5 tape read bandwidth of 140 MB/sec (per tape drive), the time cost of reading a 10 GB file is a bit over a minute, comparable in scale to the fixed cost of preparing to read, and the time cost of reading a 100 GB file is in the 10 to 15 minute range, substantially higher than the fixed cost of preparing to read. A 1 GB file, by contrast, will take on the order of 10 seconds to read, at which point the fixed cost will completely dominate the time of retrieval, wrecking overall system performance.

Note that these times don't include the time to copy the file from the PetaStore disk to the target disk, after the file is retrieved from tape. Depending on the network speed and disk speed of the target system, that time cost could vary between roughly the cost of retrieval from tape (excluding the fixed cost of preparing to read) to one or more orders of magnitude longer, in which case the time cost of retrieval from tape, even including the fixed cost, might be a modest fraction of the full retrieval cost.

These times also don't include any time that might be spent queued while waiting for resources to become available to service the retrieve request. Currently on the PetaStore, the median recall time including queue time, preparation time and read time is 1 minute 55 seconds, on an average file size of 8.4 GB, suggesting approximately 60 seconds of read time and 55 seconds of queue plus preparation time, so the average time spent in the queue is between 0 and 55 seconds (most likely closer to the low end).

Note that, in practice, the time cost of storing to the PetaStore, from the perspective of the user, is the time cost of writing to the PetaStore's disk system, because migration from disk to tape happens after the user's store request is completed.

4.6 User Training

User training typically takes roughly an hour and covers the following:

- Description and intent of the Oklahoma PetaStore
- Inquiry into the user's use case
- System rules:
  - Files MUST be 1 GB or larger.
  - Files SHOULDN'T exceed 100 GB.
  - All media must be purchased through approved channels.
- Duplication policies (and directory names)
- Interfaces (cluster, SCP/SFTP, gridFTP)
- Supplemental commands
- How to zip or tar+gzip a collection of files
- Other use case specific training as needed

5. FUTURE WORK

The PetaStore is planned to be the first of a series of large scale research data archives at OU. Bearing in mind the funding constraints of research teams – after the grant that generated a dataset ends, it can be very challenging to justify archival costs accruing to later grants, especially if the later grants don't use the older datasets that they're funding – a key goal is to minimize the later capital and/or recurring costs associated with continuing to maintain these large and growing data collections.

Taking this issue into account, the plan for the follow-on PetaStore II (whatever it may be named) is as follows:

- PetaStore II must be backward-compatible with the tape cartridge media of the current PetaStore I (though not necessarily with the software, and definitely not with the disk drive media), both in the sense of the tape format being LTO (for which generations 7 and 8 have already been announced [17]), and in the sense of the vendor of PetaStore II agreeing in writing to accept, under whatever warranty/maintenance/support contract is provided, any and all tape cartridge media from the original PetaStore I. Note that LTO tape cartridges are rated for the earliest of 15 to 30 years archival [18,19], 5000 cartridge load/unload cycles [18,20], or 200 entire-tape reads/writes [18,21]. Currently, PetaStore I tape cartridges that have ever been mounted have a mean of 29 and a median of 9 tape load/unload cycles per year, and of the 565 tape cartridges that have ever been mounted (out of 920 installed so far), only 6 (1%) have been mounted so many times that they are unlikely to last 15 years (at 5000 / 15 = 333 load/unloads per year).
- PetaStore II must support LTO tape drives that can read and write every LTO generation in the original PetaStore I, as well as at least two generations beyond PetaStore I. (Currently, the anticipated breakdown is LTO-5 and LTO-6 for PetaStore I, and LTO-7 and LTO-8 for PetaStore II). Because LTO-n tape drives (for every LTO version $n > 2$) can write to tape cartridges of LTO-n and LTO-(n-1), and can read from LTO-n, LTO-(n-1) and LTO-(n-2), and because PetaStore I currently has LTO-5, it will be necessary for PetaStore II to have at least two LTO-6 tape drives, to be able to continue to exploit PetaStore I's LTO-5 (and soon LTO-6) tape cartridges.
- The transition process from PetaStore I to PetaStore II will be as follows: Once PetaStore II is in full production, PetaStore I will be set to read-only mode, then gradually all of the files on PetaStore I will be copied to PetaStore II and deleted from PetaStore I. In particular, PetaStore II must be purchased with at least a modest quantity of (presumably LTO-7) tape cartridge media. One the first copying actions from PetaStore I to PetaStore II will copy to the new PetaStore II (LTO-7) media. Once a bulk copy has been completed and verified, those files will be deleted from PetaStore I, and as each PetaStore I tape cartridge is emptied, it will be marked ready-for-transition. Then, intermittently, collections of ready-for-transition tape cartridges will be exported from PetaStore I and imported into PetaStore II. (After completion, it may be appropriate to copy the first set of files from the new LTO-7 tape cartridge media to the old LTO-5 and/or LTO-6 media, to reduce the store and retrieve burden on these older media.)

Based on this anticipated approach, and depending on both (a) when a research team purchased a particular tape cartridge and (b) how long each PetaStore system lasts, the lifetime of a particular tape cartridge could be anywhere from 5 to 15 years, until PetaStore III (whatever form it might take). Even at the 5 year range, predictions about preferred storage solutions, based on the combination of technological progress and market conditions, are difficult, and at the 15 year range are extremely challenging. What is likely, however, is that the cost per unit of storage (e.g., per TB) will decrease substantially, anticipated in most cases to be at least one and possibly two orders of magnitude, meaning that the cost of replacing PetaStore I storage media at the onset of PetaStore III would be far lower than the original cost and thus probably realistic for many research teams.

6. ACKNOWLEDGMENTS

Portions of this material are based upon work supported by the National Science Foundation under the following grants: Grant No. OCI-1039829, "MRI: Acquisition of Extensible Petascale Storage for Data Intensive Research;" Grant no. EPS-1301789, "Adapting Socio-ecological Systems to Increased Climate Variability."

The authors are grateful to members of the OU Information Technology team for their contributions to designing and deploying the Oklahoma PetaStore, and to the OU MRI grant team for their contributions to obtaining and implementing the MRI grant.

7. REFERENCES

[1] H. Neeman, D. Brunson, J. Deaton, Z. Gray, E. Huebsch, D. Gentis and D. Horton, 2013: "The Oklahoma Cyberinfrastructure Initiative." *Proc. XSEDE 2013*.

[2] IDC, 2014: "Worldwide Smartphone Shipments Top One Billion Units for the First Time, According to IDC." http://www.idc.com/getdoc.jsp?containerId=prUS24645514 Referenced in http://www.informationweek.com/mobile/mobile-business/1-billion-smartphones-shipped-in-2013/d/d-id/1113603

[3] Wikipedia RAID webpage. http://en.wikipedia.org/wiki/RAID#Unrecoverable_read_errors_during_rebuild

[4] A. J. Argumedo, D. Berman, R. G. Biskeborn, G. Cherubini, R. D. Cideciyan, E. Eleftheriou, W. Häberle, D. J. Hellman, R. Hutchins, W. Imaino, J. Jelitto, K. Judd, P.-O. Jubert, M. A. Lantz, G. M. McClelland, T. Mittelholzer, C. Narayan, S. Ölçer and P. J. Seger, 2008: "Scaling tape-recording areal densities to 100 Gb/in^2." *IBM Journal of Research and Development,* 52 (4.5), 513-527. DOI: 10.1147/rd.524.0513.

[5] Git website. http://git-scm.com/

[6] Amazon Glacier webpage. https://aws.amazon.com/glacier/

[7] IBM System Storage DCS9900. http://www-07.ibm.com/storage/includes/content/disk/dcs/dcs9900/TSD03063USEN.pdf

[8] IBM TS3500 webpage. http://www.ibm.com/systems/storage/tape/ts3500/

[9] IBM x3650M3 webpage. http://www.ibm.com/systems/x/hardware/rack/x3650m3/

[10] IBM General Parallel File System webpage. http://www-03.ibm.com/systems/software/gpfs/

[11] IBM Tivoli Storage Manager webpage. http://www-03.ibm.com/software/products/en/tivostormana/

[12] IBM High Performance Storage System webpage. http://www4.clearlake.ibm.com/

[13] A. Cavalli, L. dell'Agnello, A. Ghiselli, D. Gregori, L. Magnoni, B. Martelli, M. Mazzucato, A. Prosperini, P. P. Ricci, E. Ronchieri, V. Sapunenko, V. Vagnoni, D. Vitlacil and R. Zappi, 2010: "SToRM-GPFS-TSM: A new approach to hierarchical storage management for the LHC experiments." *J. Phys. Conf. Ser.*, 219, 072030. http://dx.doi.org/10.1088/1742-6596/219/7/072030

[14] Shared Central File System for Research Archives (lolo Archive) webpage. http://depts.washington.edu/uwtscat/archivestorage

[15] University of Colorado PetaLibrary website. https://www.rc.colorado.edu/services/storage/petalibrary

[16] Globus Online website. https://www.globusonline.org/

[17] Ultrium LTO roadmap webpage. http://www.lto.org/media4.html

[18] Wikipedia Linear Tape Open webpage. http://en.wikipedia.org/wiki/Ultrium#Tape_durability

[19] StorageTek Linear Tape Open (LTO) Ultrium Tape Cartridges [data sheet]. Referenced in [12]. http://www.oracle.com/us/products/servers-storage/storage/tape-storage/033632.pdf

[20] StorageTek LTO Tape Cartridges - Features webpage. Referenced in [12]. http://www.oracle.com/us/products/servers-storage/storage/tape-storage/lto-ultrium-data-cartridge/features/index.html

[21] Expected Usage Life of Imation Media webpage. Referenced in [12]. http://support2.imation.com/downloads/imn/LTO/Usage_Life_Imation_Media.pdf